AMIA Annu Symp Proc 2014:1218-1227.

Desiderata for an authoritative Representation of MeSH in RDF

Rainer Winnenburg, PhD, Olivier Bodenreider, MD, PhD
National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA
{rainer.winnenburg|olivier.bodenreider}@nih.gov

The Semantic Web provides a framework for the integration of resources on the web, which facilitates information integration and interoperability. RDF is the main representation format for Linked Open Data (LOD). However, datasets are not always made available in RDF by their producers and the Semantic Web community has had to convert some of these datasets to RDF in order for these datasets to participate in the LOD cloud. As a result, the LOD cloud sometimes contains outdated, partial and even inaccurate RDF datasets. We review the LOD landscape for one of these resources, MeSH, and analyze the characteristics of six existing representations in order to identify desirable features for an authoritative version, for which we create a prototype. We illustrate the suitability of this prototype on three common use cases. NLM intends to release an authoritative representation of MeSH in RDF (beta version) in the Fall of 2014.

1 Introduction

In their seminal paper in 2001 [1], Berners-Lee et al. offer a vision of the Semantic Web featuring use cases in healthcare and the life sciences, such as accessing treatment information, finding healthcare providers and scheduling appointments. Later, Ruttenberg and members of the Health Care and Life Sciences Interest Group (HCLSIG) of the World Wide Web Consortium have highlighted the potential of the Semantic Web for supporting translational research [2].

In the era of Linked Open Data, the biomedical domain represents a significant portion of the Linked Open Data cloud [3], a growing collection of interoperable resources supported by Semantic Web technologies. As shown in Figure 1, the biomedical portion of the Linked Open Data cloud (depicted in pink) included over 40 datasets in 2011 and is still growing as datasets become available in formats suitable for Linked Data (e.g., RDF – the Resource Description Framework).

Some data providers have made their resources available in RDF (e.g., UniProt [4]). In many cases, however, the Semantic Web community has stepped up and transformed existing resources to RDF so they can participate in the Linked Open Data (LOD) cloud. According to the statistics published on the LOD cloud website [5], as of August 2011, out of the 295 datasets in the LOD cloud only 113 (39 %) were published by the data producers themselves, while 180 (61 %) were published by third-parties. For example, LinkedCT is a Linked Data version of the National Library of Medicine's registry of clinical trials, ClinicalTrials.gov, created and maintained by researchers at the University of Toronto [6], and used in several projects, including clinical registries [7].

While Linked Open Data is arguably the most visible part of the Semantic Web, Semantic Web technologies have permeated many industries, including libraries. For example, the Library of Congress has recently initiated the Bibliographic Framework Initiative (BIBFRAME) [8], an attempt to replace the legacy MARC 21 format [9] with Semantic Web technologies for the representation and exchange of bibliographic data [10]. This new framework could leverage RDF representations of legacy authority files, such as the Library of Congress Subject Headings and the National Library of Medicine's Medical Subject Headings (MeSH) [11], for the annotation of bibliographic records. However, MeSH is made available by its developer in XML, MARC, and ASCII flat files, as well as through the Unified Medical Language System (UMLS) Metathesaurus [12], but not in RDF.

This initiative prompted us to revisit our earlier attempt to produce an RDF version of MeSH, in the objective of establishing desiderata for an authoritative representation of MeSH in RDF. More specifically, we explore the Linked Open Data cloud for RDF versions of MeSH contributed by the community, including our own, and we analyze their characteristics in order to identify desirable features for an authoritative version. We propose a prototype RDF representation of MeSH that meets these criteria, and we illustrate its usefulness through three common use cases. NLM intends to release an authoritative representation of MeSH in RDF (beta version) in the Fall of 2014.

2 Background

2.1 *Semantic Web technologies*

The Semantic Web is an extension of the current Web [1, 13]. Underlying the Semantic Web are a set of technologies, including Uniform Resource Identifiers (URIs) – identifiers for resources on the Web [14], the Resource Description Framework (RDF) – a format for representing (and making statements about) Web resources [15], and the SPARQL query language for RDF repositories [16]. Ontologies provide the vocabulary and shared semantics required for annotating resources and to support inference. RDF and the Web Ontology Language (OWL) are the W3C standards for encoding data/knowledge [17]. The Simple Knowledge Organization System (SKOS) is recommended for the representation of thesauri and similar artifacts [18].

RDF describes information in the form of subject-predicate-object triples. This enables information to be represented in the form of a graph. The graph can then be queried using SPARQL. RDF has multiple serialization formats, including RDF/XML, N-Triples, Turtle and, most recently, JSON-LD. The two main distribution mechanisms for RDF data are making the RDF datasets available for download and providing a "SPARQL endpoint", i.e., a live service to which queries can be made.

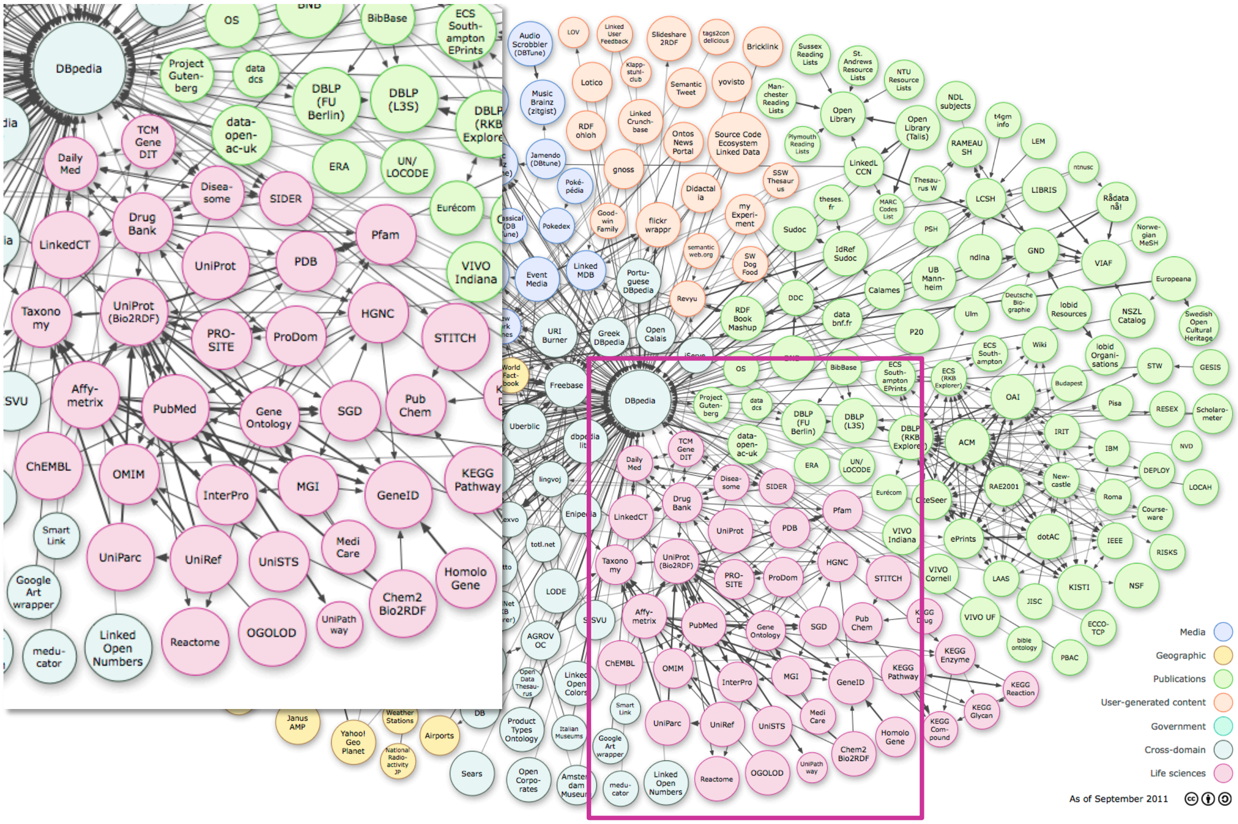

Figure 1. Linked Open Data cloud as of September 2011 (with close-up view on the life sciences portion)

2.2 *Medical Subject Headings (MeSH)*

The MeSH thesaurus is a controlled vocabulary produced by the National Library of Medicine (NLM) and used for indexing, cataloging and searching for biomedical and health-related information and documents [11]. MeSH consists of three main record types: Descriptor records, Qualifier records and Supplementary Concept records (SCRs). Each record has a unique identifier. Descriptors, also known as Main Headings, are mostly used to indicate the subject of an indexed item in NLM's MEDLINE bibliographic database and other databases. *Acquired immunodeficiency Syndrome* (D000163) is an example of a descriptor. Qualifiers, also known as subheadings, are used for indexing and cataloging in conjunction with descriptors, to indicate which specific aspect of a descriptor is discussed. For example, the qualifier *adverse effects* can be used with drug descriptors to index adverse drug events. In addition to descriptors, SCRs are used by annotators to index new or less frequently occurring terms in the literature. All SCRs are connected to at least one descriptor (“heading mapped to” in MeSH parlance). In some cases, SCRs can be mapped to multiple descriptors or to descriptor-qualifier combinations. MeSH is easily accessible via the MeSH Browser [19] and made available by NLM for download in various formats, including XML, MARC, and ASCII flat files, as well as through the Unified Medical Language System (UMLS) Metathesaurus [12].

Three features of MeSH make it non-standard. These idiosyncratic features are: a 3-level structure (descriptor / concept / term), a contextual hierarchical structure and the need in common use cases for entity combinations (descriptor-qualifier combinations) that are not materialized in MeSH.

3-level structure. Instead of the traditional concept-term terminological model of most thesauri, MeSH uses a 3-level structure. In addition to concepts and terms (i.e., concept names), MeSH also defines descriptors, i.e., small aggregates of concepts grouped together as needed to support indexing and retrieval. For example, the MeSH descriptor *Ofloxacin* (D015242) groups several concepts, including the main concept identified by M0023430 (with terms *Ofloxacin* and *Ofloxacine*), the concept for a salt of this drug, M0329515 (with term *Ofloxacin Hydrochloride*), and the concept for the experimental form of this drug before it was marketed, M0023432 (with name *Ru-43280*). This 3-level structure is not amenable to representation with standard terminological models, such as SKOS, which only accommodates concepts and terms.

Contextual hierarchical structure. In addition to a non-standard terminological model, MeSH also uses a non-standard hierarchical organization. The hierarchy among MeSH descriptors is indicated through “tree numbers” assigned to descriptors. Tree number inclusion reflects that the descriptor with the longer tree number is narrower than that with the shorter tree number. For example, the tree number for *Liver* [A03.620] has an additional node (.620) compared to that of *Digestive System* [A03], indicating the narrower relation between the two. Note that tree numbers are not the unique identifiers of descriptors. Descriptors often have multiple tree numbers reflecting particular aspects of the descriptors, each aspect being assigned specific broader and narrower descriptors. For example, the descriptor *Eye* (D005123) has two tree numbers, A01.456.505.420 and A09.371. In the A01 tree, *Eye* is narrower than *Head* [A01.456] and broader than *Eyebrows* [A01.456.505.420.338] and *Eyelids* [A01.456.505.420.504], whereas, in the A09 tree, *Eye* is narrower than *Sense Organs* [A09] and broader than *Eyelids* [A09.371.337], *Retina* [A09.371.729], *Uvea* [A09.371.894] and nine other descriptors. Note that, although *Head* is broader than *Eye* (A01 tree), some descendants of *Eye* in the A09 tree (e.g., *Retina*) do not have *Head* as their ancestor. For all practical purposes, the broader/narrower relationship among MeSH descriptors is not transitive.

Descriptor-qualifier combinations. Finally, although MeSH represents descriptors and qualifiers as separate entities, common use cases of MeSH require combinations of descriptors and qualifiers. Chief among them is MEDLINE indexing, where descriptor-qualifier combinations are assigned to articles from the biomedical literature by indexers. For example, the descriptor-qualifier combination *Levofloxacin/adverse effects* is found as an index term for this report of a dermatological adverse drug event titled “Case of drug-induced bullous pemphigoid by levofloxacin” [20]. As mentioned earlier, MeSH itself uses descriptor-qualifier combinations to relate supplementary concept records to descriptors (and qualifiers). For example, the SCR for the drug *antofloxacin* (C522674) is mapped to *Ofloxacin/analogs & derivatives*, combining the descriptor *Ofloxacin* and the qualifier *analogs & derivatives*.

3 Related work

In the absence of an authoritative representation of MeSH in RDF from the NLM, there have been several efforts over the past few years to make MeSH available for the Sematic Web, starting from various sources, making use of different transformation techniques, and adopting different schemas and models. In the following, we review these existing representations of MeSH in RDF critically, in order to examine how the developers of RDF representations of MeSH have coped with the challenges associated with representing the three idiosyncratic features of MeSH discussed earlier. References for the six sources can be found in Table 1.

3.1 Original MeSH-SKOS

In 2004, Van Assem et al. generated what is probably the first representation of MeSH in RDF, leveraging the Simple Knowledge Organization System (SKOS) RDF Schema [21]. In addition to the RDF resource itself, these researchers made available the script they had developed to create it. Although the representation they produced was essentially for proof-of-concept purposes and was never updated, interested users can still apply their transformation to more recent versions of MeSH.

Because SKOS is a concept-based model, it cannot do justice to the distinction between descriptors and concepts in MeSH. As a consequence, all terms are directly attached to the descriptor in the SKOS representation, which may constitute a limitation for some applications. Hierarchical relations among MeSH descriptors are provided through *skos:broader* relations between descriptor URIs. Descriptor-qualifier combinations are not materialized. Of note, this proof-of-concept version does not provide a complete representation of MeSH (e.g., supplementary concept records are omitted).

3.2 Science Commons MeSH-SKOS and qualified-headings

In 2006, the Van Assem transformation was slightly modified by Science Commons researchers for use in the HCLSIG 2007 demo and was part of a mashup system that seeks to help the process of bioinformatics knowledge integration [22]. Additionally, Science Commons provides a companion resource (mesh/qualified-headings), in which MeSH descriptor-qualifier combinations are materialized, e.g., *mesh:D000001Q000008 skos:prefLabel "Calcimycin - administration & dosage"*. Although the data on the Science Commons website is for MeSH 2008, the script for creating the resources is available from the website.

The main MeSH-SKOS has the limitations of representations of MeSH in SKOS discussed earlier. Although descriptor-qualifier combinations are not materialized in the main SKOS representation, they are made available as a separate resource. This feature makes it possible to refer to these combinations in the representation of MEDLINE citations developed by the same researchers, in which *has-as-major-mesh* and *has-as-minor-mesh* relations are asserted between MEDLINE records and descriptor-qualifier combinations in MeSH. This representation is interesting as it reflects a specific use case, i.e., MEDLINE indexing, where citations are often indexed with descriptor-qualifier combinations. This representation is the only one we found that treats descriptor-qualifier combinations as first-class entities, with their own URIs.

3.3 Bio2RDF MeSH

Started in 2006, the Bio2RDF project uses Semantic Web technologies to provide linked data from publicly available databases in the life sciences [23, 24]. As of early 2014, there are 28 datasets available, including MeSH, linked together with normalized URIs, and sharing a common ontology. The current representation of MeSH in Bio2RDF is derived from the original MeSH 2014 ASCII flat files, containing all descriptor, supplementary concept, and qualifier records, their relations, and metadata.

Bio2RDF MeSH does not distinguish between descriptors and concepts in MeSH and all terms are directly attached to the descriptor as literals through proprietary *entry-term* relations. Hierarchical relations among MeSH descriptors are provided through *rdfs:subClassOf* relations, while MeSH only asserts broader/narrower relations. Moreover, Bio2RDF hierarchical relations are between materialized “tree-number classes” (i.e., specific aspects of the descriptors), e.g., *mesh:A08.186.211.132.93 rdfs:subClassOf mesh:A08.186.211.132*. Descriptors are linked to these tree-number classes through *mesh_vocabulary:mesh-tree-number* relationships. Descriptor-qualifier combinations are not materialized and supplementary concept records are mapped to a literal representation of the combination instead (e.g., *mesh:C014481 mesh_vocabulary:heading-mapped-to "Codeine/*analogs & derivatives"*), which is suboptimal from a Linked Data perspective.

3.4 NCBO BioPortal UMLS-MESH

Developed by the National Center for Biomedical Ontology at Stanford University since 2006, BioPortal is an open repository of biomedical ontologies made accessible via web services and web browsers [25]. BioPortal now offers RDF versions of all its ontologies. It actually provides two versions of MeSH. The main version (MESH) is derived from various files from the Unified Medical Language System (UMLS) [12, 26] Metathesaurus distribution and is currently being used in more than ten projects (e.g., the Drug Interaction Knowledge Base). UMLS-MeSH is available for the 2014 version of MeSH.

Like most MeSH representations, the version in BioPortal does not distinguish between descriptors and concepts in MeSH. MeSH descriptors are subclasses (*rdfs:subClassOf*) of their broader descriptors according to the MeSH tree number hierarchy. The tree numbers themselves are linked to the descriptors through annotation properties. Descriptor-qualifier combinations are not materialized. Supplementary concept records (SCRs) are mapped to descriptors through *mapped_to* relations and, independently, to qualifier through *has_mapping_qualifier* relations, where applicable. Decoupling the mapping of SCRs to descriptors and qualifiers is problematic when an SCR is mapped to multiple descriptor-qualifier combinations, because there is no explicit statement of the association between descriptors and qualifiers in this case. Additionally, supplementary concept records are mapped to literals created for descriptor-qualifier combinations (e.g., *"D007830/Q000002"*), which, here again, is suboptimal from a Linked Data perspective.

3.5 NCBO BioPortal RH-MeSH

The "Robert Hoehndorf version of MeSH" (RH-MeSH) is the second version of MeSH found in BioPortal. Unlike the version of MeSH derived from the UMLS presented earlier, this version was created for a specific purpose, i.e., to facilitate the use of the descriptor and SCR hierarchy. It is used in the cross-species phenotype network, PhenomeNet. RH-MESH is represented in OWL. All MeSH descriptors and SCRs are classes as expected. Additionally, tree-numbers (representing specific aspects of descriptors) are also treated as classes and linked to other tree number classes through *rdfs:subClassOf* relationships. SCRs are represented as subclasses of descriptors. MeSH descriptors for drugs are also subclasses of the descriptors corresponding to their pharmacological actions. Except for class labels, this version of MeSH does not expose any other properties of the descriptors (e.g., definition). RH-MeSH is available for the 2014 version of MeSH.

While blurring the distinction between descriptors and SCRs (and even between descriptors and their tree numbers), this representation of MeSH provides an easy way for traversing the tree of MeSH entities. However, because it assumes that the descriptors are linked through subclass relationships, which is not what MeSH asserts, it contains inaccurate assertions (e.g., *Liver* subclass of *Digestive system*). Moreover, it links SCRs to descriptors (and not descriptor-qualifier combinations), and considers these links subclass relations (not mapping relations as indicated in MeSH), which also results in inaccurate assertions. For example, the MeSH assertion *Acrorenal Syndrome* mapped to *kidney/abnormalities* (descriptor-qualifier combination), is wrongly translated into *Acrorenal Syndrome* subclass of *kidney*.

3.6 MOR MeSH baseline

In 2009, in order to support internal research projects in the Medical Ontology Research (MOR) group at NLM, we created a fully automated process to transform the native XML representation of MeSH into RDF, based on EXtensible Stylesheet Language Transformations (XSLTs). Our goal with this representation was that it be close to the original XML representation and lossless. In other words, we made sure that all the information and only the information in the source XML had been captured during the transformation into RDF. We have updated this simple baseline representation regularly by applying our XSLTs to each new release of MeSH, but have kept both the XSLTs and RDF output internal to our research group.

The 3-level structure of MeSH with all relations between descriptors, concepts, and terms is preserved in this version. The descriptor hierarchy has been created for convenience, but kept in a separate graph, because it was not present in the native XML representation. The linkage of Supplementary Concept Records to descriptors and qualifiers is implemented through blank nodes rather than materialized descriptor-qualifier combinations, which is suboptimal as blank nodes lack shared semantics.

Table 1. Availability of existing RDF representations of MeSH

Name	Dissemination type	URL
Original MeSH-SKOS	Web site Download	http://thesauri.cs.vu.nl http://thesauri.cs.vu.nl/mesh/rdf/mesh1a.rdf
Science commons mesh-skos and qualified-headings	Web site Endpoint Download	http://neurocommons.org/page/Bundles/ http://beta.neurocommons.org http://neurocommons.org/page/Bundles/mesh/mesh-skos/ http://neurocommons.org/page/Bundles/mesh/qualified-headings/
Bio2RDF MeSH	Web site Endpoint Download	http://bio2rdf.org http://mesh.bio2rdf.org/sparql/ http://download.bio2rdf.org/release/2/mesh/
NCBO BioPortal UMLS MESH	Web site Endpoint Download	http://bioportal.bioontology.org http://sparql.bioontology.org http://bioportal.bioontology.org/ontologies/MESH/
NCBO BioPortal RH-MESH	Web site Endpoint Download	http://bioportal.bioontology.org http://sparql.bioontology.org http://bioportal.bioontology.org/ontologies/RH-MESH/
MOR MeSH baseline	Download/Endpoint	Not publicly available

4 Methods and Results

In this study we performed a review of existing representations of MeSH in RDF, established a list of desirable features for an authoritative representation, and implemented a prototype version of MeSH in RDF according to these criteria.

4.1 Analysis of the characteristics of existing RDF representations of MeSH

We conducted a manual analysis of the six existing representations of MeSH in RDF introduced in the Background section. We downloaded all representations that we reviewed and accessed them through their SPARQL endpoint whenever possible. However, we did not test any of the transformation scripts and did not create any local MeSH representations based on those.

We established a list of the characteristics of these resources, while focusing on the following features. We used the latest release date as an indication of the currency of the resource (the 2014 version of MeSH, available since September of 2013, was expected to be found). We categorized a representation as lossless only if the complete information provided in MeSH was exposed in the RDF representation. More specifically, we expected coverage of all three components of MeSH (descriptors, qualifiers and SCRs), as well as all important features (e.g., definitions). We also expected the semantics of MeSH relations to be preserved (e.g., hierarchical relations among descriptors represented as broader relations, not subclass relations). We noted the format(s) in which the resources were made available (e.g., RDF, OWL) and which specific terminological model or schema was used (e.g., SKOS). Some resources were developed as proof-of-concept and never intended to be maintained regularly, while others were in stable or beta version. We recorded this distinction. Whenever available, we added the information about the generation mechanism, as well as the original MeSH source used for creating the RDF representation (MeSH XML, MARC, or ASCII flat files, or UMLS Metathesaurus files). In terms of dissemination, we recorded whether the RDF resources were available for download or could be queried through a SPARQL endpoint, and whether the developers made the scripts used for the generation of RDF available to the community. Finally, we recorded how the idiosyncratic features of MeSH had been represented, especially hierarchical relations and descriptor-qualifier combinations.

Table 2. Characteristics of existing RDF representations of MeSH

Name	Latest Release Date	Lossless	Format	Status	Conversion	Dissemination	Features	
							Descriptor-Qualifer comb.	Hierarchical relations
Original MeSH-SKOS	2004	No	RDF (SKOS)	Proof-of-concept	Perl script, XSLT, XML	Download, Script	No	Yes (broader)
Science commons mesh-skos and qualified-headings	2008	No	RDF (SKOS)	Proof-of-concept	Using eswc06 Perl script, XSLT, XML	Endpoint, Script	Yes	Yes (broader)
Bio2RDF MeSH	2014	No	RDF	stable	PHP, ASCII	Endpoint, Download, Script	No* (as literals)	Yes (subclass)
NCBO Bioportal UMLS MESH	2014 (UMLS 2014AA)	UMLS view on MeSH	RDF	stable	UMLS Metathesaurus files	Endpoint, Download	No* (two separate relations and literals)	Yes (subclass)
NCBO Bioportal RH-MESH	2014	No	OWL	Beta	unspecified	Endpoint, Download	No	Yes (subclass)
MOR MeSH baseline	2014	Yes	RDF	Internal	XSLT	Used only internally	No* (through blank nodes)	Yes* (broader, in a separate graph)

The results of our analysis are summarized in Table 2. Four resources are up to date (i.e., reflect MeSH 2014 as of July 2014), but this was not the case at an earlier stage of our exploration a few months ago. The other two versions, developed for proof of concept, are not expected to be up to date. Most versions only capture a subset of the MeSH features, rather than all the details present in the original source. Two versions use OWL, one SKOS, and the other three use RDF with no specific terminological model. Except for one, the providers offer the resource for download, and most also provide a SPARQL endpoint. The transformation script is made available in three cases. Regarding the features of special interest to us, we found that only one resource materializes descriptor-qualifier combinations in a way that is suitable for linking to a MEDLINE dataset. While all resources investigated offer some kind of representation of hierarchical relations between MeSH descriptors, it is worth noting that the semantics of hierarchical relations had been reinterpreted by half of the providers as subclass relations, as opposed to broader relations.

4.2 Desirable features for an authoritative representation of MeSH in RDF

Based on our analysis of the characteristics of existing representations of MeSH in RDF, we compiled a list of desirable features for an authoritative representation of MeSH in RDF. Not mentioned in this list are the best practices for publishing linked data, such as guidelines for creating URIs, which are applicable to all RDF datasets, not only authoritative representations of vocabularies such as MeSH [27].

Completeness: Given the multiplicity of use cases for MeSH, it is likely that a representation of MeSH in RDF will be used in different ways by different users. Therefore, we believe it is best to provide a systematic representation in RDF of all features present in the XML version of MeSH. At a minimum, the authoritative representation of MeSH in RDF should represent those features exposed through the UMLS, as some sources do. However, representing only a subset of the MeSH entities (e.g., omitting the SCRs) would not be an option for most use cases.

Usability: As mentioned earlier, the structure of MeSH is too complex to be represented with the terminological model of SKOS. On the other hand, an RDF representation limited to the features of MeSH explicitly present in the XML version (such as our original baseline version), lacks the convenience of exposing important features, including hierarchical relations among descriptors, and materialized descriptor-qualifier combinations. Usability of the authoritative representation of MeSH in RDF should be analyzed in light of major use cases, which for MeSH include the role it plays in indexing and retrieval of the biomedical literature (MEDLINE). As illustrated by RH-MeSH, the authoritative representation could be extended by users in order to further facilitate traversal of the MeSH tree (at the expense of the original semantics of some MeSH relations).

Linkability: The authoritative representation of MeSH in RDF is meant to be linked to other resources in the Semantic Web. Although an authoritative version of MEDLINE in RDF has not been released yet, it would be a prime candidate for interoperating with MeSH in the Linked Open Data cloud. This requires coordinated development of resources within the institution developing these resources. Moreover, it requires harmonization of base URIs and predicates wherever possible. The representation provided by Science Commons, illustrated in Figure 2, prefigures what a MeSH-MEDLINE combined subset would look like.

```
@prefix c: <http://purl.org/science/owl/sciencecommons/> .
@prefix m: <http://purl.org/commons/record/mesh/> .
<http://purl.org/commons/record/pmid/11696761>
          c:has-as-minor-mesh m:D000368 ;
          c:has-as-minor-mesh m:D002292 ;
          c:has-as-minor-mesh m:D002292Q000150 ;
```

Figure 2. MeSH-MEDLINE combined subset as provided by Science Commons

Currency: A new version of MeSH becomes available each year and resources such as MEDLINE are synchronized with new versions of MeSH once a year. The authoritative representation of MeSH in RDF needs to be available in a timely fashion, and in coordination with related resources.

Availability: The authoritative representation of MeSH in RDF should be made available for download alongside the XML representation and other legacy representations. Users could load MeSH locally in an RDF database. Additionally, the resource should be available through a SPARQL endpoint so that local installation is not a requirement for use. (To some extent, this dual distribution mechanism is no different from the provision of datasets for download and web services, as is the case for the UMLS, for example.)

Transparency: Besides providing the RDF data for download, the transformation programs used to generate the RDF resource (e.g., scripts, XSLT, etc.) should be made available. This will give users insights into the transformation process and allow them to create RDF files locally, or create variants as required by their specific use cases. Exposing the transformation process might also help the community detect potential errors and suggest improvements.

4.3 Towards an authoritative representation of MeSH in RDF

The source version of MeSH we used in our prototype is the XML version (i.e., the 2014 MeSH XML files). The transformation rules from XML to RDF were coded using the XSLT (Extensible Stylesheet Language Transformations) language. The DTD file of each record type (Descriptor, Supplementary Concept Record, Qualifier) informed the creation of an XSLT for each type of MeSH record. The XSLT files were applied to the source XML files using the saxon XSLT processor. In addition to the features of MeSH explicitly represented in the XML file (and already present in our earlier baseline version), we chose to represent some other features for convenience purposes, i.e., in order to increase usability. We created a hierarchy among descriptors using the *skos:broader* relationship. We also elected to materialize descriptor-qualifier combinations for all the allowable qualifiers of each descriptor, in order to support upcoming representations of MEDLINE in RDF.

Our current prototype version of an authoritative representation of MeSH in RDF has the following features:

- Completeness: The losslessness of our transformation can easily be demonstrated by regenerating the original XML from the RDF using another set of XSLT files.
- Currency: The XSLT can easily be applied to any new version of MeSH. No changes to the XSLT are required unless changes are made to the XML DTD of the MeSH records. Because the transformation is completely automated and fast, it is conceivable to produce a nightly build reflecting the addition of supplementary concept records (e.g., for internal use by NLM for indexing purposes).
- Availability: The RDF file can easily be made available on the MeSH website, alongside the XML and legacy representations. Additionally, NLM intends to make it available through a SPARQL endpoint.
- Transparency: The XSLT files can be distributed on the MeSH website together with the XML and RDF versions.

It is difficult to comment on linkability at this prototype stage. However, the recently created NLM Linked Data Infrastructure Working Group will oversee the development of the final version of the authoritative representation of MeSH in RDF and of the other RDF datasets NLM intends to make available as Linked Data. The addition of descriptor-qualifier combinations to the prototype representation of MeSH in RDF prefigures the availability of an interoperable version of MEDLINE in RDF.

Usability is probably the most difficult criterion to fulfil and evaluate, due to the multiplicity of use cases for MeSH, beyond NLM's own use cases for indexing and retrieval of the biomedical literature. As already mentioned, while a complete, lossless version may best serve some complex use cases, other, more common use cases may be best served by simpler representations. Input from the community will help NLM determine which trade-offs to adopt for the final representation.

5 Discussion

In the following we provide several use cases illustrating the application of our prototype authoritative representation of MeSH in RDF. We also discuss current limitations and future directions for our work.

5.1 Use cases for an authoritative representation of MeSH in RDF

Currency: The Bibliographic Framework Initiative (BIBFRAME) [8, 10] initiated by the Library of Congress (LOC) provides a model for expressing and connecting bibliographic data on the web to replace the current standard for bibliographic exchange (MARC 21) [9]. BIBFRAME catalogues works and their instances, and associates them with authorities, which are resources that represent persons, organizations, topics, etc., (e.g., LOC subject headings can be accessed through the URI namespace <http://id.loc.gov/authorities/subjects/>). MeSH, as an authoritative resource, should be used for providing authorities for biomedical subjects (such as diseases, treatments, etc.). URIs for descriptors and SCRs should be added to catalog records where applicable. In BIBFRAME, topics can be assigned in combination with publication types. An extension of the representation of MeSH in RDF – possibly local to a given cataloging site – could include descriptor-publication type combinations (similar to the descriptor-qualifier combinations included in our prototype).

Quality assurance of MeSH: One of the key advantages of linked data is the possibility to link information across different data sources. But even for a single, locally available RDF graph, SPARQL queries can help gather information from the data, which would be difficult to retrieve using other representations (e.g. flat files, XML). We helped the MeSH development team detect and remove cyclic relationships in the MeSH graph and assess that the 2014 version of MeSH is acyclic. Other applications include summarizing all supplementary concept records for a given pharmacologic action.

Linked data applications: In a recent collaboration with the FDA Center for Drug Evaluation and Research (CDER), we developed a novel analytic tool for quantitative drug-adverse event (ADE) safety signal detection based on mining the biomedical literature (MEDLINE). We leveraged the MeSH indexing terms to extract associations between co-occurring drug entities (in the context of adverse effects) and clinical manifestations (induced by chemicals). Information about ADEs is captured by different kinds of entities in MeSH (main headings, pharmacological actions, and supplementary concept records) and their inter-relations. In addition to the representation of MeSH in RDF, we created a prototype version of MEDLINE in RDF for the subset of articles under investigation in our study. The use of Semantic Web technologies enabled us to perform complex queries across the MeSH and MEDLINE datasets and greatly facilitated our work.

5.2 Limitations and future work

This prototype of an authoritative representation of MeSH in RDF is current and complete, but not final. Important improvements have been made already to our original baseline version, including the explicit representation of the hierarchical relations among descriptors (to improve usability) and the materialization of descriptor-qualifier combinations (to improve linkability with MEDLINE). However, additional editorial changes have to be made. For example, base URIs, namespaces and predicate names were chosen somewhat arbitrarily in the early development phase, where the focus was on demonstrating feasibility and scalability of the transformation method. However, these elements become important as this prototype is evolving into the authoritative representation of MeSH in RDF. Feedback from the user community will also inform future developments.

In order to accommodate some features of the XML version, our baseline version had introduced blank nodes to represent, for example, entry combinations and concept relations. While developing the current prototype, we critically reexamined earlier design choices and found that in most cases blank nodes were unnecessary or could be replaced by materialized combinations (e.g., for descriptors and qualifiers). All the blank nodes created initially were removed.

6 Conclusions

In the absence of an authoritative version of MeSH in RDF from the NLM, there have been several efforts over the past few years to make MeSH available for the Sematic Web. We identified six existing representations of MeSH in RDF and conducted a manual analysis of these representations. Based on the characteristics of these resources we compiled a list of desirable features for an authoritative representation of MeSH in RDF (completeness, usability, linkability, currency, availability and transparency). We implemented a prototype of an authoritative representation of MeSH in RDF that fulfills these criteria and illustrated its suitability on three use cases. Our prototype was influenced by our early baseline representation in RDF of all features present in the XML version of MeSH. However, we made substantial changes in order to improve usability and linkability. NLM intends to release an authoritative representation of MeSH in RDF in the Fall of 2014. We believe that the availability of such a resource will foster the adoption of MeSH in biomedical Semantic Web applications.

Acknowledgements

This work was supported by the Intramural Research Program of the NIH, National Library of Medicine (NLM). The authors would like to thank former NLM colleagues Ramez Gazzaoui and Genaro Hernandez, who contributed to an early prototype of MeSH in RDF in 2009, and Nancy Fallgren for sharing her insights on the BIBFRAME project. Thanks to the NLM Linked Data Infrastructure Working Group and to colleagues from the MeSH development team for their encouragement and support.

References

1. Berners-Lee T, Hendler J, Lassila O. The Semantic Web - A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Sci Am. 2001 May;284(5):34-+.
2. Ruttenberg A, Clark T, Bug W, Samwald M, Bodenreider O, Chen H, et al. Advancing translational research with the Semantic Web. BMC Bioinformatics. 2007;8 Suppl 3:S2.
3. Cyganiak R, Jentzsch A. LOD Cloud. Available from: http://lod-cloud.net/.
4. Redaschi N, Consortium U. UniProt in RDF: Tackling Data Integration and Distributed Annotation with the Semantic Web.: Nature Precedings; 2009; Available from: http://precedings.nature.com/documents/3193/version/1.
5. Biebl M, Hakaim AG, Hugl B, Oldenburg WA, Paz-Fumagalli R, McKinney JM, et al. Endovascular aortic aneurysm repair with the Zenith AAA Endovascular Graft: does gender affect procedural success, postoperative morbidity, or early survival? Am Surg. 2005 Dec;71(12):1001-8.
6. Hassanzadeh O, Kementsietsidis A, Lim L, Miller RJ, Wang M. LinkedCT: A Linked Data Space for Clinical Trials. CoRR. 2009;abs/0908.0567.
7. da Silva KR, Costa R, Crevelari ES, Lacerda MS, de Moraes Albertini CM, Filho MM, et al. Glocal clinical registries: pacemaker registry design and implementation for global and local integration--methodology and case study. PLoS One. 2013;8(7):e71090.
8. Pathak J, Kiefer RC, Bielinski SJ, Chute CG. Applying semantic web technologies for phenome-wide scan using an electronic health record linked Biobank. J Biomed Semantics. 2012;3(1):10.
9. MARC 21 format for bibliographic data 1999 Edition Update No. 17. Library of Congress; 2013 [cited 2014 March 10]; Available from: http://www.loc.gov/marc/bibliographic/.
10. Miller E, Ogbuji U, Mueller V, MacDougall K. Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Services. Washington, DC: Library of Congress 2012 November 21.
11. Nelson SJ, D. JW, L. HB. Relationships in Medical Subject Headings (MeSH). In: Bean CA, Green R, editors. Relationships in the organization of knowledge. Dordrecht; Boston: Kluwer Academics Publishers; 2001. p. 171-84.
12. NLM. Unified Medical Language System (UMLS). 2013; Available from: https://uts.nlm.nih.gov/.
13. Tse LW, Steinmetz OK, Abraham CZ, Valenti DA, Mackenzie KS, Obrand DI, et al. Branched endovascular stent-graft for suprarenal aortic aneurysm: the future of aortic stent-grafting? Can J Surg. 2004 Aug;47(4):257-62.
14. W3C. URI. Available from: http://www.w3.org/TR/uri-clarification/.
15. W3C. RDF. Available from: http://www.w3.org/RDF/.
16. W3C. SPARQL. Available from: http://www.w3.org/TR/sparql11-overview/.
17. W3C. OWL. Available from: http://www.w3.org/TR/owl2-overview/.
18. W3C. SKOS. Available from: http://www.w3.org/2004/02/skos/.
19. NLM. MeSH Browser. 2014; Available from: https://www.nlm.nih.gov/mesh/MBrowser.html.
20. Ma HJ, Hu R, Jia CY, Yang Y, Song LJ. Case of drug-induced bullous pemphigoid by levofloxacin. J Dermatol. 2012 Dec;39(12):1086-7.
21. van Assem M, Menken MR, Schreiber G, Wielemaker J, Wielinga B, editors. A Method for Converting Thesauri to RDF/OWL. 3rd Int'l Semantic Web Conf (ISWC'04); 2004: Springer-Verlag.
22. Kanda J, Kaynar L, Kanda Y, Prasad VK, Parikh SH, Lan L, et al. Pre-engraftment syndrome after myeloablative dual umbilical cord blood transplantation: risk factors and response to treatment. Bone Marrow Transplant. 2013 Jan 21.
23. Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. J Biomed Inform. 2008 Oct;41(5):706-16.
24. Callahan A, Cruz-Toledo J, Dumontier M. Ontology-Based Querying with Bio2RDF's Linked Open Data. J Biomed Semantics. 2013 Apr 15;4 Suppl 1:S1.
25. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W170-3.
26. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D267-70.
27. W3C. Best Practices for Publishing Linked Data.   [updated 2014]; Available from: http://www.w3.org/TR/ld-bp/.